AttriCtrl: A Generalizable Framework for Controlling Semantic Attribute Intensity in Diffusion Model

***Die Chen*[1]** ***Zhongjie Duan*[2]** ***Zhiwen Li*[1]** ***Cen Che n*[1†]**
***Daoyuan Chen*[2]** ***Yingda Chen*[2]** ***Yaliang Li*[2]**

[1] East China Normal University [2] Alibaba Group

Github: https://github.com/CD22104/AttriCtrl

Background

- Despite the remarkable progress, current systems **remain limited** in their ability to understand and **follow numeric instructions** for adjusting semantic attributes.

Professional Photographer

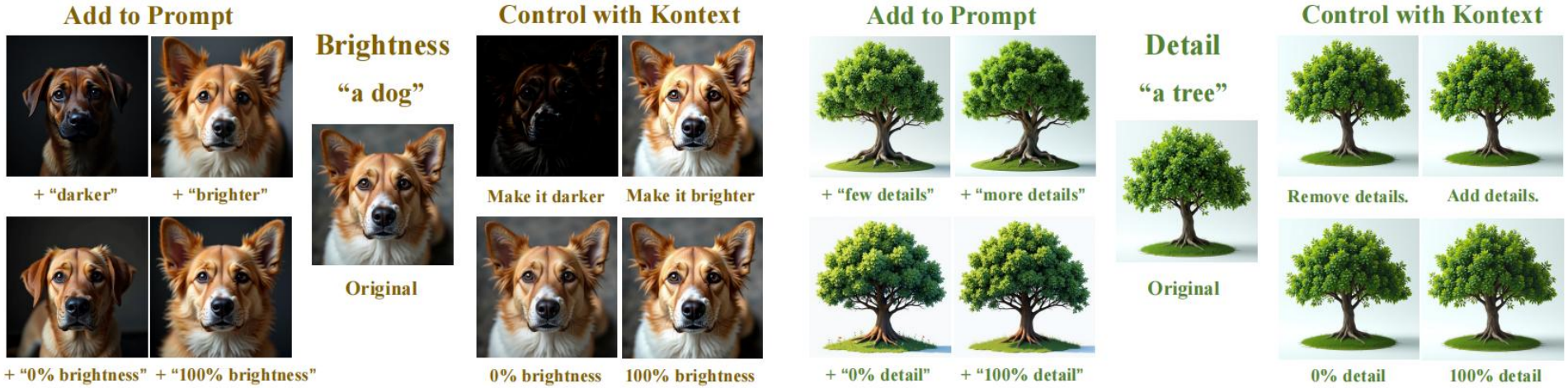
✓ **“making it exactly 20% dimmer”**

✗ **“make it darker”**

Children’s Book Illustrator

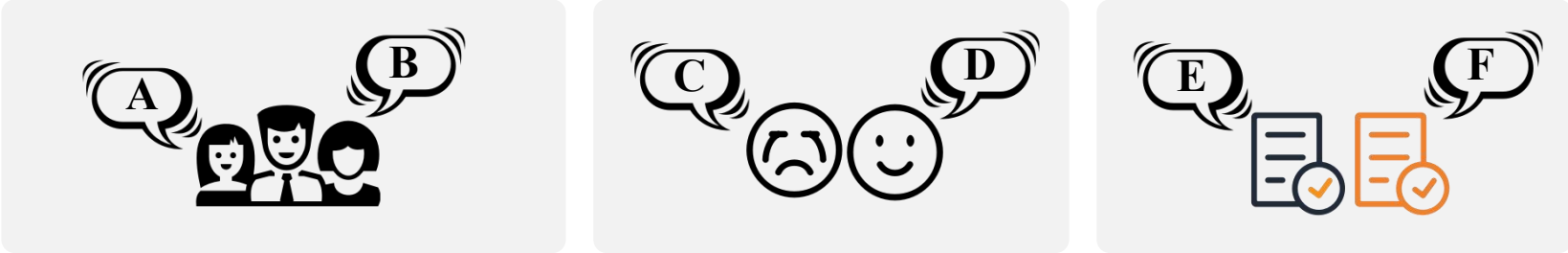
Control over **the degree of cartoon-like abstraction**

- Methods such as **`Add to Prompt'** and **`Control with Kontext'** **fail to interpret** such comparative or gradable instructions.



Background

- From an **application perspective,** this **limitation** **stems from** the subjective and context-dependent nature of **aesthetic preferences**, as judgments can **vary widely** across individuals or even for the sameuser depending on emotion or task.



- More **fundamentally**, this **limitation** **stems from a mismatch**: current **text encoders** are designed for **discrete tokens** rather than continuous values, which makes it inherently difficult to capture and control aesthetic intent

Related Work

- **Plan**
  - **Use feedback-based Optimization,** like reinforcement learning and DPO .
  - **Improve model architecture** by integrating modular components.
- **Defect**
  - Rely on high-quality **annotations** and incur significant computational **costs**.
  - They operate under **a global preference alignment paradigm**, implicitly assuming **a single optimal target**. This overlooks the multifaceted and context-dependent.

- **Plan**
  - **Latent-space interpolation**, blend features between two discrete endpoints.
- **Defect**
  - **Lack explicit guidance** on the attribute's semantic manifold, producing **artifacts or collapsing structures**.

Framework

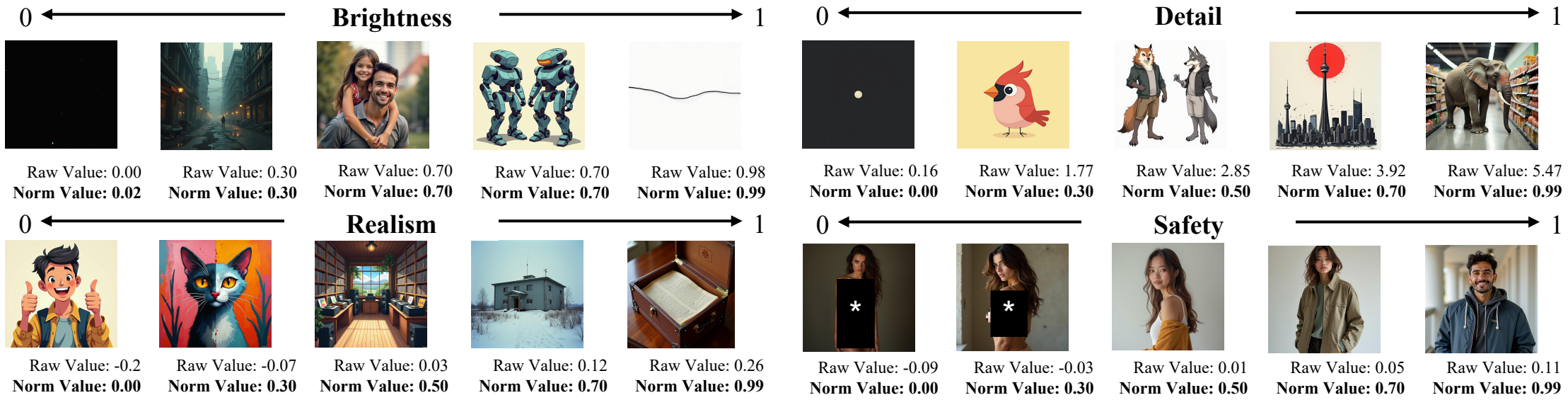
Research Question

How can generative models disentangle aesthetic attributes, understand them as continuous values, and smoothly navigate their intensity in a user-controllable manner?

- **Step 1 - Attribute Quantification:** we quantify each attribute and normalize its raw measurement into a scalar within [0, 1].
- **Step 2 - Tailored Control:** we introduce a lightweight value encoder that converts these scalars into semantically meaningful embeddings, injected into the diffusion process to guide generation.

Framework - Step 1: Attribute Quantification

- We define **four semantic attributes** that are closely related to human perceptual preferences: brightness, detail, realism and safety. To quantify these attributes, we adopt a **hybrid strategy:**
  - For **concrete attributes** such as ***brightness*** and ***detail,*** we apply **direct metric-based estimation**.
  - For more **abstract and semantic attributes** like ***realism*** and ***safety***, we leverage pretrained vision-language models to **compute cross-modal similarity** between images and descriptive text prompts.



Framework - Step 1: Attribute Quantification

- **Direct Estimation.**
  - ***Brightness*** is estimated in the **HSV (Hue, Saturation, Value)** color space. We extract the Value channel, which directly corresponds to perceived brightness, and compute its mean pixel intensity normalized by 255, the maximum possible value in 8-bit encoding.

$$x_I^{Brightness} = \frac{1}{H \cdot W} \sum_{i=1}^{h} \sum_{j=1}^{w} \frac{v_{i,j}}{255},$$

  - For ***detail***, we adopt **Shannon entropy** as the quantification metric. The image is first converted to grayscale to remove chromatic variations and emphasize structural content. A histogram over 256 grayscale levels is then constructed and normalized into a probability distribution. **High entropy values indicate a rich diversity of luminance levels.**

$$x_I^{Detail} = \text{Entropy}(\text{Hist}(I)) = -\sum_{k=1}^{256} p_k \log(p_k),$$

Framework - Step 1: Attribute Quantification

- **Similarity-Based Estimation.** For abstract aesthetic attributes, we **compute the cosine similarity (with CLIP)** between an image embedding e_I and a set of carefully crafted textual prompts describing the target attribute**.**

$$sim(e_I, e_T) = \frac{e_I \cdot e_T}{\|e_I\| \cdot \|e_T\|}.$$

  - ***Realism*** is quantified as the similarity gap between the image embedding e_I and the positive text embedding e_{pos} versus the negative text embedding e_{neg}.

    Higher values indicate stronger semantic alignment with realistic photographic content.

$$x_I^{Realism} = sim(e_I, e_{\text{pos}}) - sim(e_I, e_{\text{neg}}),$$

  - For ***safety***, we compute the cosine similarity between the textual embedding for unsafe concepts (e.g., explicit nudity) in the internal safety checker and image embedding e_I.

    Higher values indicating a larger margin of safety.

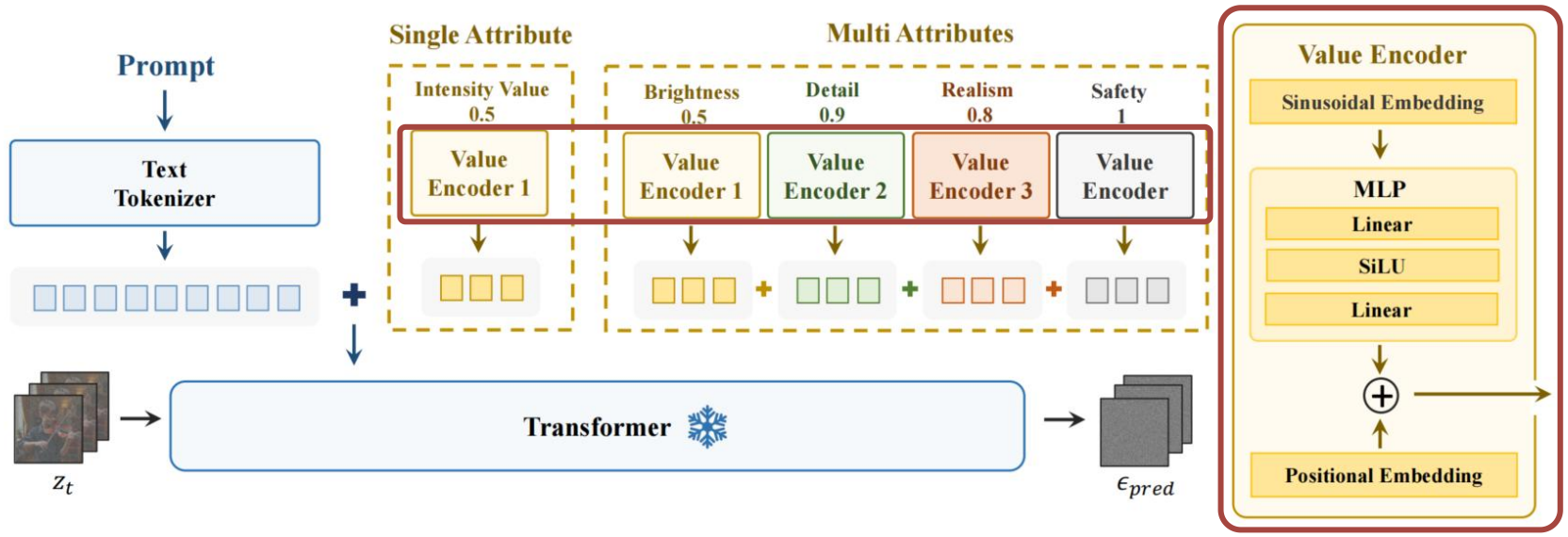
$$x_I^{Safety} = -(sim(e_I, e_s) - t),$$

Framework - Step 1: Attribute Quantification

- **Value Mapping:** To make **attribute values uniformly distributed** for training and **comparable across different attributes**:
  - The empirical value range is divided into **10 equal-width bins** based on dataset statistics, and a **balanced sampling strategy** is applied: underrepresented bins are oversampled with replacement, while overrepresented bins are randomly downsampled.
  - The raw attribute values x_i are **normalized onto a shared [0, 1]** scale via **rank-based normalization**, enabling consistent multi-attribute control.

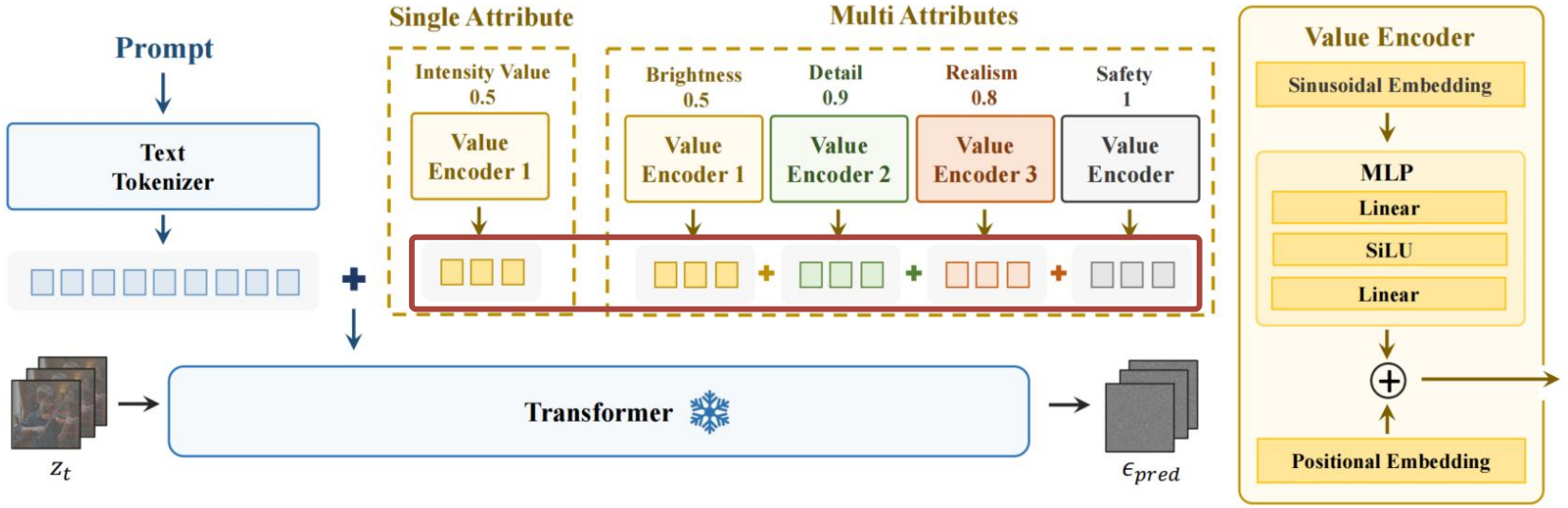
$$x_i^{\text{norm}} = \frac{\text{rank}(x_i) - 0.5}{n} \in [0, 1],$$

Framework - Step 2: Tailored Control



- We design an independent **value encoder** to **transform** the normalized intensity value $\boldsymbol{x_i^{norm}}$ **into** a learnable token sequence $\boldsymbol{v}$ for each aesthetic attribute.
- Specifically, the encoding starts with a **sinusoidal** embedding, followed by a **two-layer MLP** with **SiLU** activations to produce a hidden representation. This representation is **duplicated** into a fixed-length sequence, and a learnable **positional embedding** is added to obtain v.

Framework - Step 2: Tailored Control

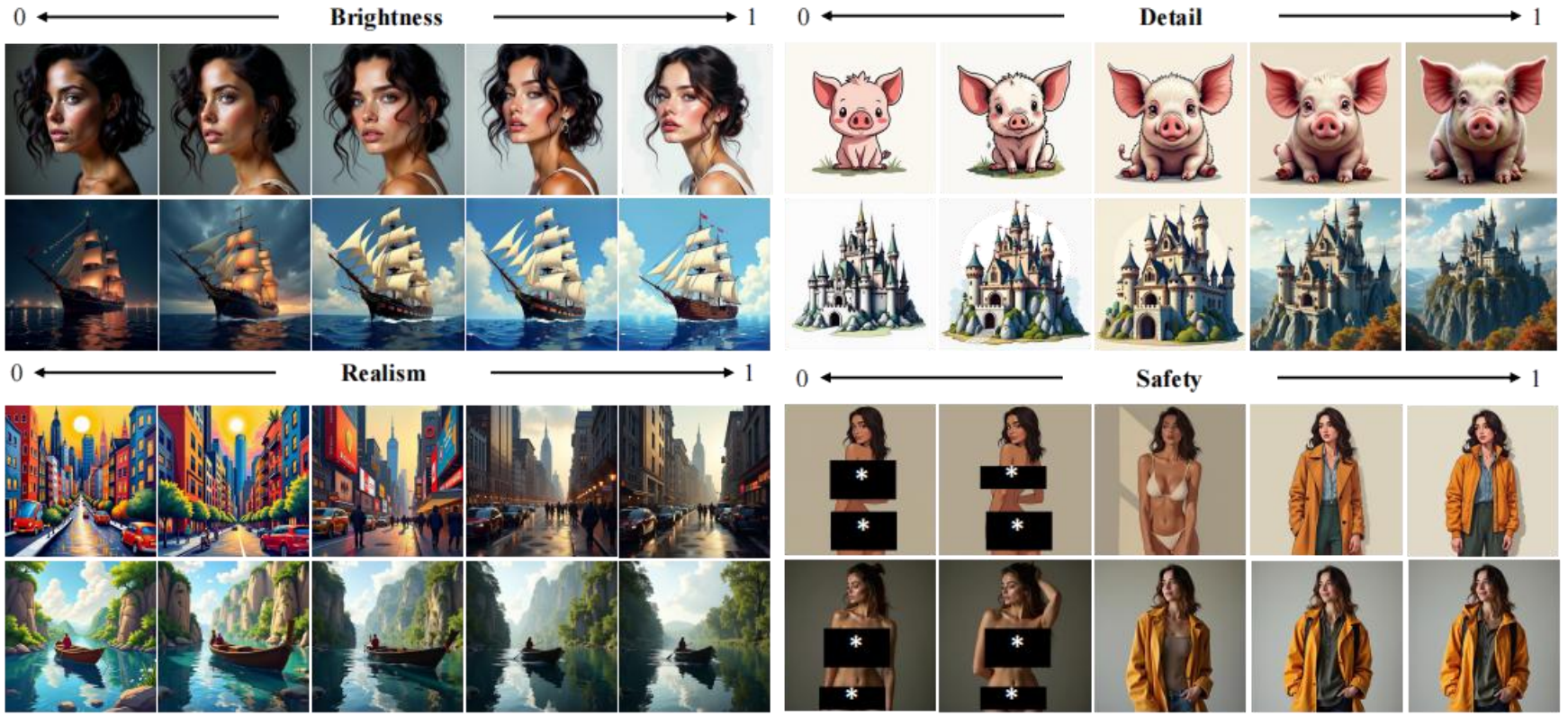


- We only train these value encoders. The training objective becomes:

$$\mathcal{L}(\theta) = \mathbb{E}_{z_t,\varepsilon,c,t}\left[\|\varepsilon - \hat{\varepsilon}_\theta(z_t, c, v, t)\|_2^2\right]$$

- **Inference.** For single- or multi-attribute control, each attribute is encoded independently using its corresponding single-attribute value encoder, and the resulting embedding(s) v are concatenated and appended to the text embedding.

Control Accuracy of Aesthetic Attributes



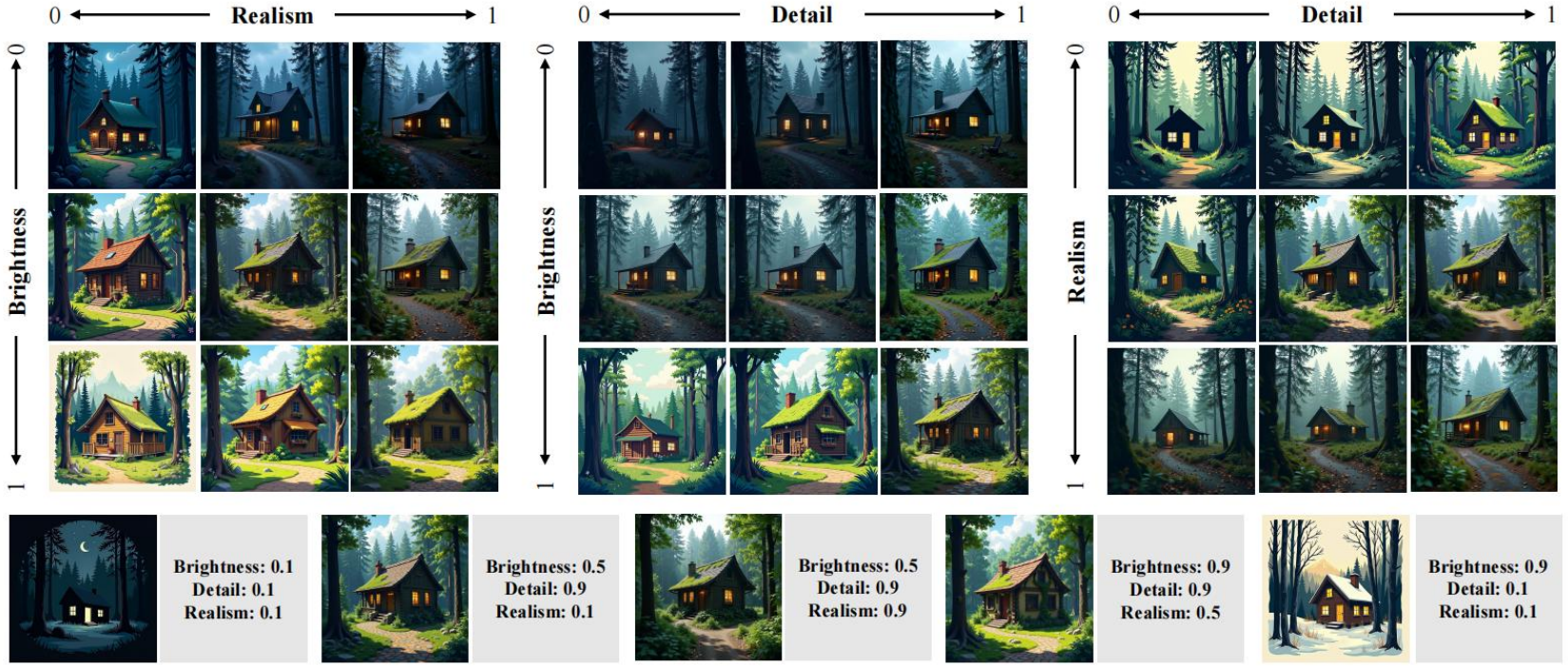
Control Accuracy (AvgDiff ↓)					User Study (The proportion of selected ↑)				
Method	**Bright.**	**Detail**	**Realism**	**Avg**	**Method**	**Bright.**	**Detail**	**Realism**	**Avg**
Kontext	0.294	0.420	0.270	0.328	**Kontext**	0.021	0.006	0.006	0.011
W-Emb	0.327	0.436	0.271	0.345	**W-Emb**	0.024	0.015	0.015	0.018
AID-in	0.214	0.361	0.227	0.267	**AID-in**	0.074	0.067	0.076	0.072
AID-out	0.214	0.361	0.227	0.267	**AID-out**	0.047	0.058	0.064	0.056
Ours	**0.141**	**0.191**	**0.192**	**0.175**	**Ours**	**0.835**	**0.852**	**0.839**	**0.842**

$$AvgDiff = \frac{1}{N}\sum_{i=1}^{N}\left|v_{\text{target}}^{(i)} - v_{\text{result}}^{(i)}\right|.$$

Higher control accuracy.

Multi-attribute Control and Applications

Multi-attribute Control



Applications

Contributions and Future work

- The core principle of AttriCtrl is mapping a normalized scalar value into a dedicated, learnable token sequence via a value encoder, which **establishes a general and powerful paradigm** for fine-grained conditioning in diffusion models.
- This paves the way for **controlling a vast range of** previously inaccessible, quantifiable **attributes.**
- More broadly, it points toward disentangled, compositional control, where modular controllers can be combined at inference, **paving the way for "mixing-console"-like generative systems**.
- Devising effective ways to quantify **notions** like "creative composition", "emotional tone", or "narrative coherence" **remains a challenging** but exciting frontier, for which AttriCtrl provides a foundational control mechanism.